Capítulo 11
Análisis exploratorio
El procedimiento *Explorar*

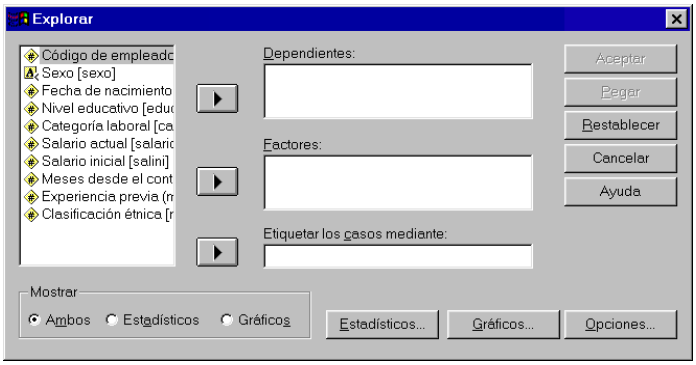
Independientemente de la complejidad de los datos disponibles y del procedimiento estadístico que se tenga intención de utilizar, una exploración minuciosa de los datos previa al inicio de cualquier análisis posee importantes ventajas que un analista de datos no puede pasar por alto. Una exploración minuciosa de los datos permite identificar, entre otras cosas: posibles errores (datos mal introducidos, respuestas mal codificadas, etc.), valores extremos (valores que se alejan demasiado del resto), pautas extrañas en los datos (valores que se repiten demasiado o que no aparecen nunca, etc.), variabilidad no esperada (demasiados casos en una de las dos colas de la distribución, demasiada concentración en torno a determinado valor), etc. El procedimiento **Explorar** permite estudiar este tipo de problemas.

Explorar

Además de incluir gran parte de los estadísticos descriptivos ya estudiados en los procedimientos **Frecuencias** y **Descriptivos**, el procedimiento **Explorar** permite obtener nuevos estadísticos descriptivos, identificar casos atípicos y estudiar con mayor precisión la forma y otras características de una distribución. También permite contrastar dos de los supuestos en que se basan muchas de las técnicas de análisis que estudiaremos más adelante: normalidad y homogeneidad de varianzas. Para obtener todos estos estadísticos:

- Seleccionar la opción **Estadísticos descriptivos** > **Explorar** del menú **Analizar** para acceder al cuadro de diálogo *Explorar* que muestra la figura 11.1.

Figura 11.1. Cuadro de diálogo *Explorar*.



Dependientes. Trasladando a la lista **Dependientes** una o más variables de la lista de variables del archivo y pulsando el botón **Aceptar** obtenemos los estadísticos y gráficos que el procedimiento **Explorar** ofrece por defecto: varios estadísticos descriptivos, un diagrama de tallo y hojas y un diagrama de caja.

Factores. Si en lugar de un análisis referido a todos los casos del archivo, deseamos análisis separados para diferentes grupos de casos (por ejemplo, para hombres y para mujeres, o para cada categoría laboral, etc.), podemos introducir la variable que define esos grupos en la lista **Factores**. Si introducimos más de una variable *factor*, obtendremos, para cada variable *dependiente*, un análisis completo referido a cada uno de los grupos definidos por cada variable *factor*. (Mediante sintaxis, es posible obtener información referida a subgrupos definidos por la combinación de dos o más *factores*).

Etiquetar los casos mediante. En los diagramas de cajas y en los resultados que incluyen listados de casos, los casos individuales son identificados por el número de registro (fila) que ocupan en el *editor de datos*. Si deseamos utilizar como identificadores de caso los valores de alguna variable, no tenemos más que introducir esa variable en este cuadro.

Mostrar. Las opciones de este recuadro permiten seleccionar qué parte del análisis deseamos obtener: estadísticos y gráficos, sólo estadísticos o sólo gráficos.

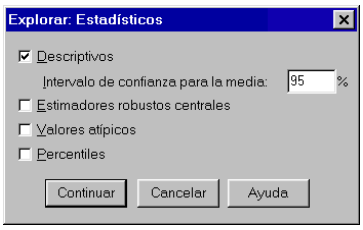
Estadísticos

La opción **Estadísticos** (ver figura 11.1) permite obtener algunos estadísticos adicionales a los que ofrece el procedimiento **Explorar** por defecto. Para seleccionar estos estadísticos:

- Pulsar el botón **Estadísticos...** del cuadro de diálogo *Explorar* (ver figura 11.1) para acceder al subcuadro de diálogo *Explorar: Estadísticos* que muestra la figura 11.2.

  Para que el botón **Estadísticos...** esté activo es necesario que en el recuadro **Mostrar** esté marcada la opción **Ambos** o la opción **Estadísticos**.

Figura 11.2. Subcuadro de diálogo *Explorar: Estadísticos*.



- **Descriptivos**. Esta opción, que se encuentra activa por defecto, ofrece la media aritmética, la mediana, la media truncada o recortada al 5 % (media aritmética calculada eliminando el 5 % de los casos con valores más pequeños y el 5 % de los casos con valores más grandes con el objetivo obtener una media menos sensible a la presencia de valores extremos),

el intervalo de confianza para la media, el error típico de la media, la varianza, la desviación típica, el valor mínimo, el valor máximo, la amplitud, la amplitud intercuartílica, los índices de asimetría y curtosis, y los errores típicos de los índices de asimetría y curtosis.

Intervalo de confianza para la media: *k* %. Permite fijar el nivel de confianza con el que deseamos obtener el intervalo de confianza para la media. El valor de *k* por defecto es 95, pero es posible introducir cualquier otro valor entre 1 y 99,99.

☐ **Estimadores robustos centrales**. Son estimadores de tendencia central basados en el método de máxima verosimilitud (de ahí que también sean conocidos como estimadores-M). En realidad, un estimador robusto central o estimador-M no es más que una media ponderada en la que los pesos asignados a los casos dependen de la distancia de cada caso al centro de la distribución: los casos centrales reciben un peso de 1 y los demás valores reciben un peso tanto menor cuanto más alejados se encuentran del centro.

Al igual que ocurre con la media truncada, los estimadores-M son menos sensibles que la media aritmética a la presencia de valores extremos. Por tanto, cuando las distribuciones son muy asimétricas, es preferible utilizar como índices de tendencia central, en lugar de la media aritmética, los estimadores-M.

Existen varios estimadores-M que difieren entre sí por la forma concreta de asignar pesos a los casos. El procedimiento **Explorar** incluye cuatro de estos estimadores: Huber, Andrew, Hampel y Tukey (puede encontrarse una descripción detallada de estos estimadores en Norusis, 1993, págs. 192-194; y en Palmer, 1999, págs. 125-162).

☐ **Valores atípicos**. Muestra los 5 casos con valores más pequeños y los cinco casos con valores más grandes. Si existen empates en los valores ocupados por el quinto caso más pequeño o el quinto más grande, la salida muestra un mensaje indicado tal circunstancia. (Mediante sintaxis, puede controlarse el número de casos atípicos listados).

☐ **Percentiles**. Muestra los percentiles 5, 10, 25, 50, 75, 90 y 95. El SPSS incluye diferentes métodos para calcular percentiles. Marcando esta opción se obtienen percentiles calculados con el método HAVERAGE, que consiste en asignar al percentil buscado el valor que ocupa la posición $i = p\,(n + 1)$ cuando los casos están ordenados de forma ascendente; p se refiere a la proporción de casos que acumula el percentil buscado (por ejemplo, el percentil 30 acumula una proporción de casos de 0,30), y n se refiere al tamaño de la muestra). Si el valor de i no es un número entero, el valor del percentil se obtiene por interpolación: $X_i(1 - d) + X_{i+1}(d)$, donde X_i se refiere al valor que ocupa la posición correspondiente a la parte entera de i, y d se refiere a la parte decimal de i.

El SPSS incluye (disponibles mediante sintaxis) otros cuatro métodos de cálculo de percentiles:

- El método WAVEREAGE es idéntico en todo al método HAVERAGE excepto en un detalle: $i = np$.
- El método ROUND asigna al percentil buscado el valor que ocupa la posición correspondiente a la parte entera de $i = np + 0{,}5$ (o, lo que es lo mismo, la posición entera más próxima a $i = np$).

- El método EMPIRICAL asigna el valor que ocupa la posición $i = np$ cuando i es un número entero, y el valor que ocupa la posición siguiente a la parte entera de i cuando $i = np$ es un número decimal.
- El método AEMPIRICAL asigna la media de X_i y X_{i+1} cuando $i = np$ es un número entero, y asigna el valor que ocupa la posición siguiente a la parte entera de i cuando $i = np$ es un número decimal.

Los resultados también muestran las *bisagras* de Tukey: una versión distinta de los clásicos cuartiles. La primera bisagra (similar al percentil 25) es el valor que ocupa la posición intermedia entre la mediana y el valor más pequeño de la distribución; la segunda bisagra es la mediana; la tercera bisagra (similar al percentil 75) es el valor que ocupa la posición intermedia entre la mediana y el valor más grande de la distribución.

Ejemplo (Estadísticos descriptivos > Explorar > Estadísticos)

Este ejemplo muestra cómo obtener los estadísticos del procedimiento **Explorar**. Vamos a utilizar para ello la variable *salario* como *dependiente* y la variable *sexo* como *factor*.

- En el cuadro de diálogo *Explorar* (ver figura 11.1), trasladar la variable *salario actual* a la lista **Dependientes** y la variable *sexo del empleado* a la lista **Factores**.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Explorar: Estadísticos* (ver figura 11.2).
- En el subcuadro de diálogo *Explorar: Estadísticos*, marcar todas las opciones: **Descriptivos**, **Estimadores robustos centrales**, **Valores atípicos** y **Percentiles**.

Aceptando estas elecciones, el *Visor* ofrece varias tablas con la información solicitada.

Tabla 11.1. Tabla de *Descriptivos* del procedimiento *Explorar.*

Salario actual

		Sexo			
		Hombre		Mujer	
		Estadístico	Error típ.	Estadístico	Error típ.
Media		$41,441.78	$1,213.97	$26,031.92	$514.26
Intervalo de confianza para la media al 95%	Límite inferior	$39,051.19		$25,018.29	
	Límite superior	$43,832.37		$27,045.55	
Media recortada al 5%		$39,445.87		$25,248.30	
Mediana		$32,850.00		$24,300.00	
Varianza		380219336,303		57123688,268	
Desv. típ.		$19,499.21		$7,558.02	
Mínimo		$19,650		$15,750	
Máximo		$135,000		$58,125	
Rango		$115,350		$42,375	
Amplitud intercuartil		$22,675.00		$7,012.50	
Asimetría		1,639	,152	1,863	,166
Curtosis		2,780	,302	4,641	,330

La tabla 11.1 muestra los estadísticos generados por la opción **Descriptivos**. Se trata de los estadísticos descriptivos clásicos: media, mediana, varianza, desviación típica, rango, índices de asimetría y curtosis, etc. Como novedad, encontramos dos estadísticos descriptivos no recogidos en los procedimientos **Frecuencias** y **Descriptivos**: la media *recortada* al 5 por ciento, que adopta un valor más cercano a la mediana que a la media aritmética (lo que demuestra que es menos sensible que ésta a la presencia de valores extremos); y la amplitud intercuartílica, que refleja la distancia existente entre los cuartiles 1 y 3. La salida ofrece también los límites del intervalo de confianza para la media al 95 por ciento: podemos estimar, con una confianza del 95 por ciento, que el salario medio verdadero de los *hombres* se encuentra entre 39.051,19 y 43.832,37 dólares.

Tabla 11.2. Tabla de *Estimadores -M* del procedimiento *Explorar*.

Salario actual

Sexo	Estimador-M de Huber	Biponderado de Tukey	Estimador-M de Hampel	Onda de Andrews
Hombre	$34,820.15	$31,779.76	$34,020.57	$31,732.27
Mujer	$24,607.10	$24,014.73	$24,421.16	$24,004.51

La tabla 11.2 recoge los estimadores robustos centrales o estimadores-M. Todos ellos oscilan en torno a 33.000 $ en el grupo de varones y en torno a 24.000 $ en el grupo de mujeres, lo cual representa una estimación de la tendencia central más próxima a los valores de la mediana y de la media recortada que a los de la media aritmética (lo que revela, de nuevo, la sensibilidad de la media aritmética a la presencia de valores extremos).

Tabla 11.3. Tabla *Percentiles* del procedimiento *Explorar*.

Salario actual

	Sexo			
	Hombre		Mujer	
Percentiles	Promedio ponderado	Bisagras de Tukey	Promedio ponderado	Bisagras de Tukey
5	$23,212.50		$16,950.00	
10	$25,500.00		$18,660.00	
25	$28,050.00	$28,050.00	$21,487.50	$21,525.00
50	$32,850.00	$32,850.00	$24,300.00	$24,300.00
75	$50,725.00	$50,550.00	$28,500.00	$28,500.00
90	$69,325.00		$34,890.00	
95	$81,312.50		$40,912.50	

La tabla 11.3. muestra los percentiles 5, 10, 25, 50, 75, 90 y 95 calculados con el método *Waverage*. Si nos detenemos a estudiar el valor de los percentiles, descubriremos que nos encontramos ante una distribución asimétrica positiva: la distancia entre el percentil 10 y el 50 es de 2.200 $, mientras que la distancia entre el percentil 50 y el 90 es de 6.216 $ (más del triple).

Junto con los percentiles, aparecen las tres *bisagras* de Tukey. En el grupo de *hombres*, la tercera bisagra difiere ligeramente del tercer cuartil (percentil 75) calculado con el método *Waverage*; en el grupo de mujeres, la primera bisagra difiere ligeramente del primer cuartil (percentil 25).

Tabla 11.4. Tabla *Valores extremos* del procedimiento *Explorar*.

Salario actual

	Sexo							
	Hombre				Mujer			
	Mayores		Menores		Mayores		Menores	
	Número del caso	Valor	Número del caso	Valor	Número del caso	Valor	Número del caso	Valor
1	29	$135,000	192	$19,650	371	$58,125	378	$15,750
2	32	$110,625	258	$21,300	348	$56,750	338	$15,900
3	18	$103,750	372	$21,300	468	$55,750	411	$16,200
4	343	$103,500	22	$21,750	240	$54,375	224	$16,200
5	446	$100,000	65	$21,900	72	$54,000	90	$16,200

La tabla 11.4 ofrece un listado de los 10 casos con valores más extremos: los cinco casos con los valores más pequeños y los cinco casos con los valores más grandes. Vemos, por ejemplo, que por la parte alta de la distribución de los hombres, hay cinco casos con salarios de 100.000 $ o más, mientras que en la distribución de las mujeres los casos con mayor salario no llegan a 60.000 $.

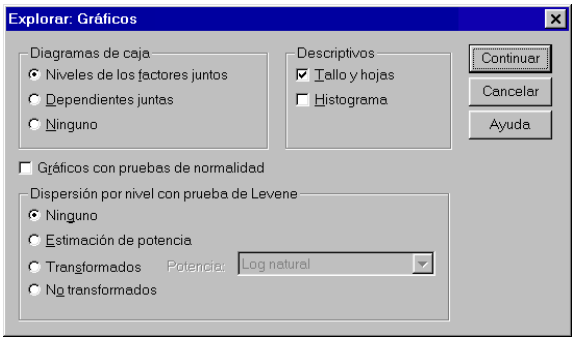
Gráficos

La opción **Gráficos** (ver figura 11.1) ofrece la posibilidad de obtener varios tipos de gráficos (diagramas de caja, diagramas de tallo y hojas, histogramas, gráficos de normalidad y de dispersión) y algunos estadísticos relacionados con los supuestos de normalidad y homogeneidad de varianzas. Para obtener estos gráficos y estadísticos:

- Pulsar el botón **Gráficos...** del cuadro de diálogo *Explorar* (ver figura 11.1) para acceder al subcuadro de diálogo *Explorar: Gráficos* que muestra la figura 11.3.

  El botón **Gráficos...** sólo se encuentra activo si en el recuadro **Mostrar** está marcada la opción **Ambos** o la opción **Gráficos**.

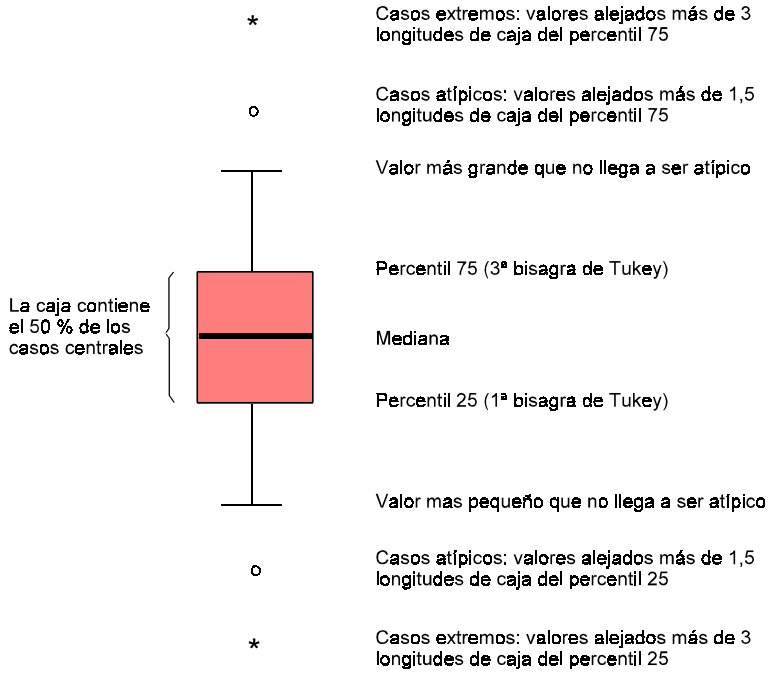
Figura 11.3. Subcuadro de diálogo *Explorar: Gráficos*.



Diagramas de caja

La figura 11.4 describe los detalles de un diagrama de caja. El diagrama incluye la mediana, los percentiles 25 y 75 (en realidad son las bisagras de Tukey), y una serie de valores (atípicos, extremos) que junto con la mediana y la propia caja proporcionan información bastante completa sobre, entre otras cosas, el *grado de dispersión* de los datos y el *grado de asimetría* de la distribución (ver Tukey, 1977).

Figura 11.4. Destalles de un diagrama de caja.



Las opciones del recuadro **Diagramas de caja** (ver figura 11.3) permiten decidir si deseamos o no obtener diagramas de caja y, en caso afirmativo, optar entre dos formas diferentes de organizar los diagramas solicitados:

- **Niveles de los factores juntos**. Muestra un gráfico diferente para cada variable dependiente. En cada uno de esos gráficos aparecen juntos los diagramas de caja correspondientes a los grupos definidos por una variable *factor*. Si no se ha seleccionado ninguna variable *factor*, cada gráfico muestra un solo diagrama de caja: el correspondiente a toda la muestra. Esta opción resulta útil para comparar distintos grupos en la misma variable. Es la opción por defecto.

- En el cuadro de diálogo *Explorar* (ver figura 11.1), trasladar las variables *salario inicial* y *salario actual* a la lista **Dependientes** y la variable *sexo del empleado* a la lista **Factores**.
- Pulsar el botón **Gráficos...** para acceder al subcuadro de diálogo *Explorar: Gráficos* (ver figura 11.3) y marcar la opción **Niveles de los factores juntos** del recuadro **Diagramas de caja**.

Aceptando estas elecciones, obtenemos los diagramas de caja que muestran las figuras 11.5.a y 11.5.b.

Figura 11.5.a. Diagramas de caja de la variable *salini* (salario inicial) en *hombres* y *mujeres*.

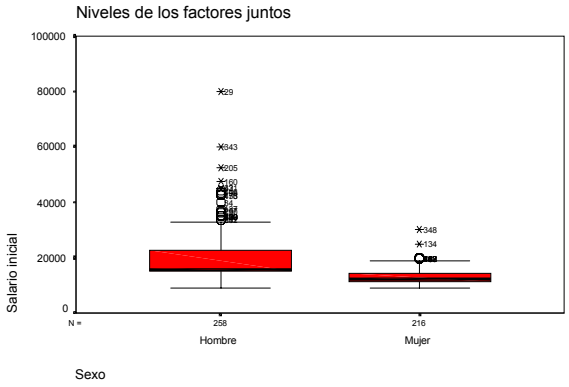
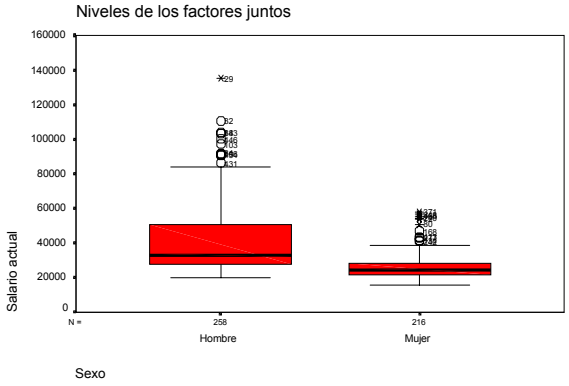


Figura 11.5.b. Diagrama de caja de la variable *salario* (salario actual) en *hombres* y *mujeres*.



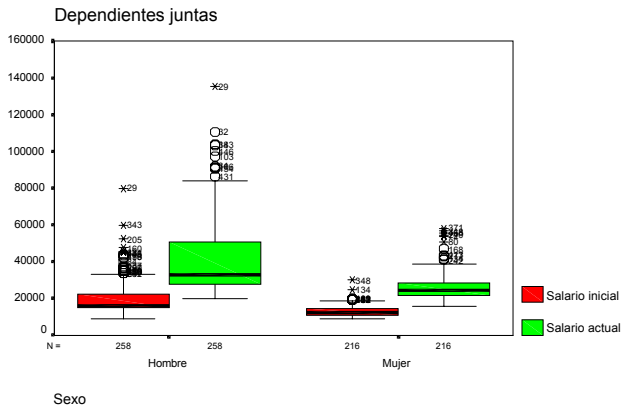
Vemos, en primer lugar, que la opción **Niveles de los factores juntos** genera dos gráficos: uno para cada variable dependiente seleccionada. En cada gráfico aparecen juntos los diagramas de caja correspondientes a los grupos definidos por la variable *factor* (*sexo* en este caso). Las medianas nos informan del salario medio de cada grupo: la mediana de los *hombres* es algo mayor que la de las *mujeres* en ambas variables dependientes. Una mediana desplazada del centro de la caja delata la presencia de asimetría: de nuevo en ambas va-

riables, la mediana del grupo de *hombres* está desplazada hacia abajo, lo que indica asimetría positiva. Las cajas (cuya altura representa la amplitud intercuartílica) muestran el grado de dispersión del 50 % de los casos centrales: en ambas variables, las cajas correspondientes al grupo de *hombres* reflejan una amplitud mayor que las cajas del grupo de mujeres. Los bigotes y los casos atípicos y extremos indican hacia dónde se desplazan los valores más alejados del centro. En todos los casos se observan valores extremos (asteriscos) y atípicos (círculos) por la parte alta de las distribuciones, lo cual indica, de nuevo, asimetría positiva (de forma más acusada en el grupo de *hombres*).

○ **Dependientes juntas**. Muestra un gráfico diferente para cada grupo de los definidos por la variable *factor*. En cada uno de esos gráficos aparecen juntos los diagramas de caja correspondientes a cada variable *dependiente*. Si el número de variables dependientes y de grupos es pequeño, el SPSS intenta mostrar en un solo gráfico todos los diagramas solicitados. Si no se ha seleccionado ninguna variable *factor*, aparece un solo gráfico con tantos diagramas de caja como variables *dependientes*. Esta opción resulta útil para comparar diferentes variables dentro del mismo grupo.

▣ Seleccionar las mismas variables dependientes (*salini* y *salario*) y la misma variable factor (*sexo*), pero marcar en el recuadro **Diagramas de caja** la opción **Dependientes juntas**. Obtenemos así el diagrama que muestra la figura 11.6.

Figura 11.6. Diagramas de caja de las variables *salini* (salario inicial) y *salario* (salario actual) en cada nivel de la variable *sexo*.



Ahora aparecen juntos los diagramas de caja correspondientes a cada variable dependiente. Como el número de variables dependientes y de factores es pequeño, en lugar de generar dos gráficos, uno para cada grupo, el SPSS ha construido un solo gráfico con los dos grupos. Con estos diagramas podemos extraer las mismas conclusiones que con los diagramas de las figuras 11.5.a y 11.5.b. Pero ahora, además, constatamos que los promedios de la variable *salario actual* son más altos que los de la variable *salario inicial*.

○ **Ninguno**. Suprime de los resultados los diagramas de caja.

Diagramas Descriptivos

Este recuadro incluye dos tipos de gráficos descriptivos: *diagramas de tallo y hoja*s e *histogramas*:

- **Tallo y hojas**. Esta opción, que se encuentra activa por defecto, permite obtener gráficos similares a los histogramas, pero que proporcionan información más precisa que éstos (ver Tukey, 1977). La figura 11.7 muestra el *diagrama de tallo y hojas* de la variable *edad* (ver apéndice) obtenido con una muestra de 148 sujetos.

Figura 11.7. Diagrama de tallo y hojas de la variable *edad*.

```
Frequency    Stem &  Leaf
   12.00        2 s  666677777777
   28.00        2 .  8888888888888888888899999999
   30.00        3 *  00000000000000011111111111111
   13.00        3 t  2222222222333
    7.00        3 f  4445555
    4.00        3 s  6666
   12.00        3 .  888899999999
   15.00        4 *  000011111111111
    4.00        4 t  3333
    4.00        4 f  4444
    7.00        4 s  6666666
    4.00        4 .  8889
    3.00        5 *  011
    3.00        5 t  222
    2.00 Extremes    (59) (64)
Stem width:     10.00
Each leaf:       1 case(s)
```

Al igual que en un histograma, la longitud de las líneas refleja el número de casos que pertenecen a cada intervalo. Cada caso (o grupo de casos) está representado por un número que coincide con el valor de ese caso en la variable. En un diagrama de tallo y hojas cada valor se descompone en dos partes: el primer o primeros dígitos (el tallo o *stem*) y el dígito que sigue a los utilizados en el tallo (las hojas o *leaf*). Por ejemplo, el valor 23 puede descomponerse en un tallo de 2 y una hoja de 3; el valor 12.300 puede descomponerse en un tallo de 12 y una hoja de 3; etc.

Cada tallo puede ocupar una o más filas. Si un tallo ocupa una sola fila, sus hojas contienen dígitos del 0 al 9. Si ocupa dos filas, las hojas de la primera fila contienen dígitos del 0 al 4 y las de la segunda fila dígitos del 5 al 9. Etc. En el diagrama de la figura 11.7, los tallos (excepto el primero y el último) ocupan cinco filas: la primera fila contiene los dígitos 0 y 1 (encabezadas con un asterisco); la segunda, los dígitos 2 y 3 (*t* = *two*, *three*); la tercera, los dígitos 4 y 5 (*f* = *four*, *five*); la cuarta, los dígitos 6 y 7 (*s* = *six*, *seven*); y la quinta, los dígitos 8 y 9 (encabezadas con un punto).

El ancho del tallo viene indicado en la parte inferior del diagrama (*stem width*) y es un dato imprescindible para interpretar correctamente el diagrama. En el ejemplo de la figura 11.7, el tallo tiene un ancho de 10, lo que significa que los valores del tallo hay que multiplicarlos por 10. Así, un tallo de 1 vale 10, un tallo de 2 vale 20, un tallo de 5 vale 50, etc.

Las hojas completan la información al tallo. Así, un tallo de 4 con una hoja de 3 representa una edad de 43 años; un tallo de 5 con una hoja de 0 representa una edad de 50 años. Etc. El número de casos que representa cada hoja (cada hoja puede representar a más de un caso) viene indicado en *each leaf*. Así, en la figura 11.7: *each leaf* = 1 caso; en la figura 11.8: *each leaf* = 3 casos.

Cuando el ancho del tallo vale 10 (como en el diagrama de la figura 11.7), los dígitos de las hojas son unidades; cuando el ancho del tallo vale 100, los dígitos de las hojas son decenas; cuando el ancho del tallo vale 1.000 (como en el diagrama de la figura 11.8), los dígitos de las hojas son centenas. Etc.

La figura 11.8 muestra el diagrama de tallo y hojas de la variable *salario inicial* (*salini*). El ancho del tallo ahora vale 1.000, lo que indica que un tallo de 1 equivale a 1.000, un tallo de 2 equivale a 2.000, un tallo de 12 equivale a 12.000, etc. Así, en el tallo 10 hay 7(3) = 21 casos cuyo salario es de 10.200 $, 1(3) = 3 casos cuyo salario es 10.500 $, etc.

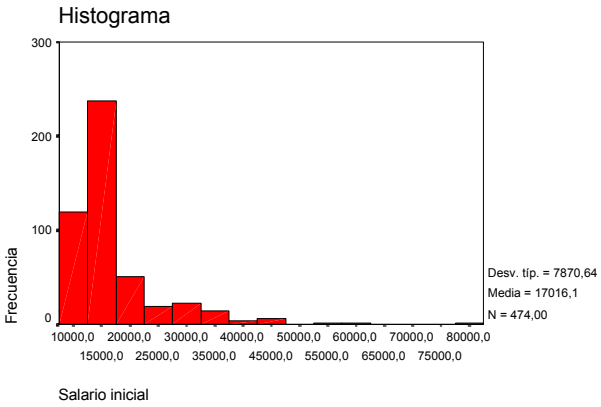
Figura 11.8. Diagrama de tallo y hojas de la variable *salario inicial*.

```
Frequency    Stem &  Leaf

    11,00        9 .  077
    42,00       10 .  2222222599999&
    40,00       11 .  12222222224555
    43,00       12 .  0000004477777&
    46,00       13 .  025555555555899&
    30,00       14 .  1222222245&
   104,00       15 .  00000000000000003777777777777777&
    29,00       16 .  0555555555&
    11,00       17 .  2224&
    20,00       18 .  0000077
    14,00       19 .  55559
     4,00       20 .  2&
    13,00       21 .  027&
     1,00       22 .  &
     4,00       23 .  27
     1,00       24 .  &
     1,00       25 .  &
    60,00 Extremes    (>=25500)

Stem width:      1000
Each leaf:       3 case(s)
```

La última fila del diagrama muestra el número de casos con valores extremos y los valores concretos que toman esos casos (entre paréntesis). Así, por ejemplo, en el diagrama de la figura 11.7 aparecen 2 casos extremos, con edades de 59 y 64 años; y en el diagrama de la figura 11.8, 60 casos extremos con un salario de al menos 25.500 $.

- **Histograma**. Un histograma se construye agrupando los datos en intervalos de la misma amplitud y levantando barras de altura proporcional al número de casos de cada intervalo. Aunque esta opción permite obtener histogramas con amplitud calculada de forma automática, tanto la amplitud de los intervalos y como otros aspectos del histograma pueden controlarse utilizando el *Editor de gráficos*. La figura 11.9 muestra un histograma de la variable *salini*. Se trata de la misma variable representada en el diagrama de tallo y hojas de la figura 11.8, por lo que podemos comparar ambos diagramas y observar las coincidencias y diferencias existentes entre ellos.

Figura 11.9. Histograma de la variable *salario inicial*.



Cómo contrastar supuestos

Muchos de los procedimientos estadísticos que estudiaremos en los próximos capítulos se apoyan en dos supuestos básicos: 1) *normalidad*: las muestras con las que trabajamos proceden de poblaciones normalmente distribuidas, y 2) *homocedasticidad* u *homogeneidad de varianzas*: todas esas poblaciones normales poseen la misma varianza. El subcuadro de diálogo *Explorar: Gráficos* (figura 11.3) incluye varios estadísticos y gráficos para contrastar estos supuestos.

Normalidad

- **Gráficos con pruebas de normalidad**. Esta opción permite obtener dos gráficos de normalidad (*Q-Q normal* y *Q-Q normal sin tendencia*) junto con dos pruebas de significación: *Kolmogorov-Smirnov* (Kolmogorov, 1933; Smirnov, 1948; Lillieffors, 1967) y *Shapiro-Wilk* (Shapiro y Wilk, 1965).

  Las pruebas de significación permiten contrastar la hipótesis de que las muestras obtenidas proceden de poblaciones normales. El SPSS ofrece, por defecto, el estadístico de Kolmogorov-Smirnov (con las probabilidades de Lillieffors para el caso en el que la media y la varianza poblacionales son desconocidas y necesitan ser estimadas). Y sólo en el caso de que el tamaño muestral sea igual o menor que 50 ofrece además el estadístico de Shapiro y Wilk. Para obtener estos estadísticos:

  - En el cuadro de diálogo *Explorar* (ver figura 11.1), trasladar la variable *salario actual* a la lista **Dependientes** y la variable *nivel de estudios* (ver apéndice) a la lista **Factores**.
  - Pulsar el botón **Gráficos...** (ver figura 11.1) para acceder al subcuadro de diálogo *Explorar: Gráficos* (ver figura 11.3), y marcar la opción **Gráficos con pruebas de normalidad** para obtener el resultado que muestra la tabla 11.5.

Tabla 11.5. Tabla *Prueba de normalidad* del procedimiento *Explorar*.

Salario actual

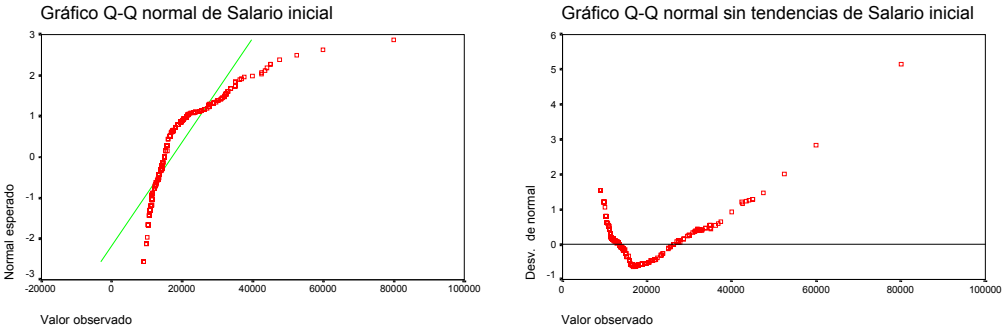
		Nivel de estudios			
		Primarios	Secundarios	Medios	Superiores
Kolmogorov-Smirnov[a]	Estadístico	,119	,079	,154	,113
	gl	53	190	181	50
	Sig.	,057	,006	,000	,148
Shapiro-Wilk	Estadístico				,951
	gl				50
	Sig.				,076

a. Corrección de la significación de Lilliefors

La tabla 11.5 ofrece los estadísticos de Kolmogorov-Smirnov y de Shapiro-Wilk acompañados de sus correspondientes niveles críticos (*Sig.* = *Significación*). Ambos permiten contrastar la hipótesis nula de que los datos muestrales proceden de poblaciones normales. Rechazaremos la hipótesis de normalidad cuando el nivel crítico (*Sig.*) sea menor que el nivel de significación establecido (generalmente 0,05). En el ejemplo, sólo los estadísticos del grupo *primarios* y del grupo *superiores* tienen asociados niveles críticos mayores que 0,05, lo que debe llevarnos a concluir que el salario de los grupos *secundarios* y *medios* no procede de poblaciones normales. El estadístico de Shapiro-Wilk sólo aparece con el nivel de estudios *superiores* porque es el único grupo con un tamaño igual o menor que 50.

El problema de estos y otros estadísticos de normalidad es que, con muestras muy grandes, son demasiado sensibles a pequeñas desviaciones de la normalidad. Por esta razón, es conveniente acompañar estos estadísticos con algún gráfico de normalidad. Según hemos señalado ya, el procedimiento **Explorar** ofrece dos gráficos de normalidad: el *Q-Q normal* y el *Q-Q normal sin tendencia* (ver figura 11.10).

Figura 11.10. Gráficos de normalidad del procedimiento *Explorar*.



En un gráfico *Q-Q normal*, cada valor observado (Y_i) es comparado con la puntuación típica NZ_i que teóricamente le correspondería a ese valor en una distribución normal estandarizada (para comprender cómo se calculan esas puntuaciones típicas normales puede consultarse, en el capítulo 5, el apartado *Asignar rangos: Tipos de rangos*). En el eje de abscisas están representados los valores observados ordenados desde el más pequeño al más grande (Y_i); en el de ordenadas están representadas las puntuaciones típicas normales (NZ_i). Cuando una muestra

procede de una población normal, los puntos correspondientes a cada par se encuentran agrupados en torno a la diagonal representada en el diagrama. Las desviaciones de la diagonal indican desviaciones de la normalidad.

Un gráfico *Q-Q normal sin tendencia* muestra las *diferencias* existentes entre la puntuación típica observada de cada valor (Z_i) y su correspondiente puntuación típica normal (NZ_i). Es decir, muestra las distancias verticales existentes entre cada punto del gráfico *Q-Q normal* y la recta diagonal. En el eje de abscisas están representados los valores observados (Y_i) y en el de ordenadas el tamaño de las diferencias entre las puntuaciones típicas observadas y las esperadas (Z_i–NZ_i). Si la muestra procede de una población normal, esas diferencias deben oscilar de forma aleatoria en torno al valor cero (línea recta horizontal). La presencia de pautas de variación no aleatorias indica desviaciones de la normalidad.

Los ejemplos de las figuras 11.11.*a*, 11.11.*b* y 11.11.*c* pueden ayudarnos a comprender el significado de los gráficos de normalidad. Estos diagramas recogen el comportamiento de tres muestras de puntuaciones aleatoriamente extraídas de tres distribuciones de probabilidad diferentes: una distribución *normal*, una distribución *uniforme* y una distribución *ji-cuadrado*.

Figura 11.11.a. Gráficos de normalidad: muestra extraída de una distribución *normal*.

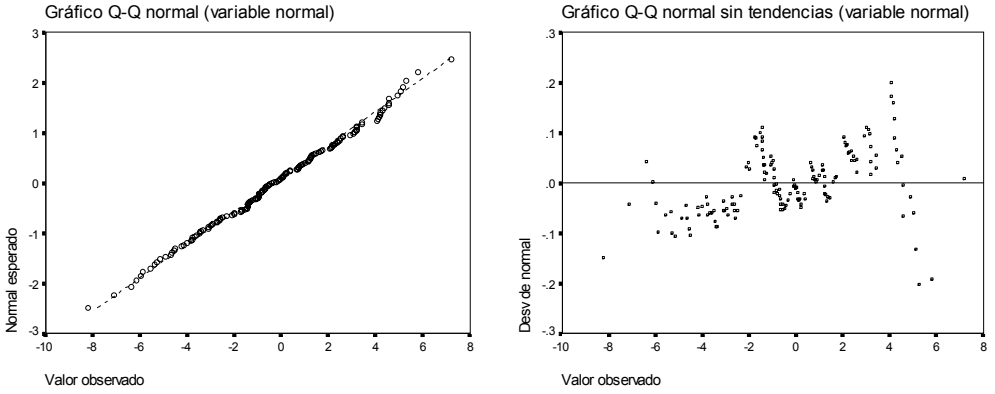


Figura 11.11.b. Gráficos de normalidad: muestra extraída de una distribución *uniforme*.

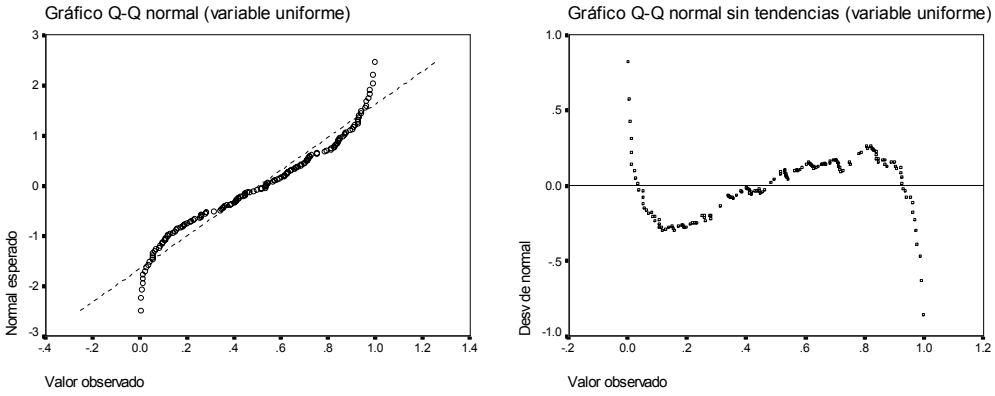
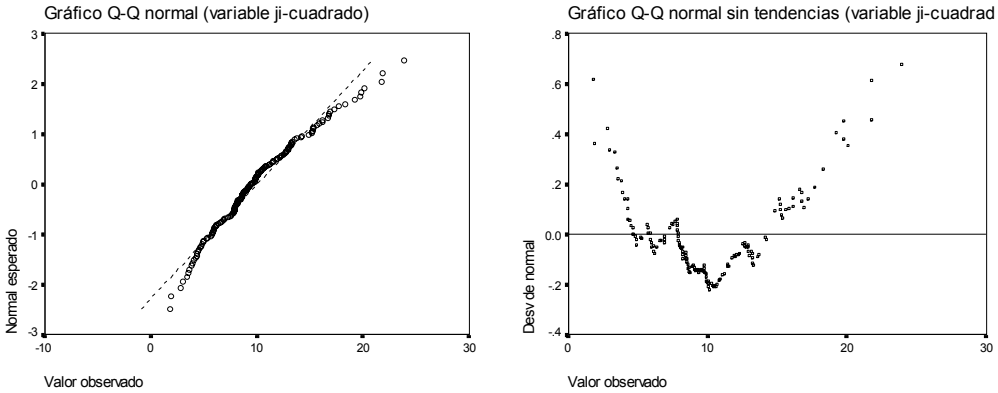


Figura 11.11.c. Gráficos de normalidad: muestra extraída de una distribución *ji-cuadrado*.



Podemos observar que, cuando una muestra de puntuaciones se distribuye normalmente (figura 11.11.*a*), los puntos del diagrama *Q-Q normal* se ajustan a la diagonal y los puntos del diagrama *Q-Q normal sin tendencia* se distribuyen aleatoriamente sin mostrar una pauta clara. Por el contrario, cuando una muestra de puntuaciones procede de una distribución uniforme (figura 11.11.*b*) o de una distribución *ji*-cuadrado (figura 11.11.*c*), los puntos del diagrama *Q-Q normal* no se ajustan a la diagonal y los puntos del diagrama *Q-Q normal sin tendencia* muestran una pauta de variación claramente no aleatoria.

Homogeneidad de varianzas

Además del supuesto de normalidad recién estudiado, el procedimiento **Explorar** también permite contrastar el supuesto de homogeneidad de varianzas (lo que requiere, obviamente, haber seleccionado al menos una variable en la lista **Factor** del cuadro de diálogo *Explorar* –ver figura 11.1).

El recuadro **Dispersión por nivel con prueba de Levene** del subcuadro de diálogo *Explorar: Gráficos* (ver figura 11.3) proporciona: 1) la prueba de *Levene* (1960) para contrastar la hipótesis de que los grupos definidos por la variable *factor* proceden de poblaciones con la misma varianza y 2) un gráfico de dispersión de la variable dependiente en cada nivel definido por la variable *factor* (gráfico de *dispersión por nivel*).

La ***prueba de Levene*** consiste en llevar a cabo una *Análisis de varianza de un factor* (ver capítulo 14) utilizando como variable dependiente la diferencia en valor absoluto entre cada puntuación individual y la media (o la mediana, o la media recortada) de su grupo. Para obtener la prueba de Levene:

- En el cuadro de diálogo *Explorar* (ver figura 11.1), trasladar la variable *salario actual* a la lista **Dependientes** y la variable *nivel de estudios* (ver apéndice) a la lista **Factores**.
- Pulsar el botón **Gráficos** para acceder al cuadro de diálogo *Explorar: Gráficos* (ver figura 11.3) y marcar la opción **No transformados**.

Aceptando estas elecciones, obtenemos el estadístico de Levene que recoge la tabla 11.6. Para obtener el estadístico de Levene es necesario que en el recuadro **Mostrar** del cuadro de diálogo *Explorar* (ver figura 11.1) esté marcada la opción **Ambos**.

Tabla 11.6. Tabla *Prueba de homogeneidad de varianzas* del procedimiento *Explorar*.

Salario actual

	Estadístico de Levene	gl1	gl2	Sig.
Basándose en la media	28,085	3	470	,000
Basándose en la mediana.	21,799	3	470	,000
Basándose en la mediana y con gl corregido	21,799	3	266,893	,000
Basándose en la media recortada	24,767	3	470	,000

El nivel crítico (*Sig.*) asociado al estadístico de Levene permite contrastar la hipótesis de homogeneidad de varianzas: si el valor del nivel crítico es menor que 0,05, podemos rechazar la hipótesis de homogeneidad. En el ejemplo, el nivel crítico (cualquiera que sea el estimador de tendencia central a partir del cual obtengamos las diferencias) vale 0,000, por lo que podemos afirmar que la varianza de la variable *salario actual* no es la misma en las cuatro poblaciones definidas por la variable *estudios*.

Además de la prueba de Levene, el recuadro **Dispersión por nivel con prueba de Levene** (figura 11.3) contiene varias opciones relacionadas con el ***gráfico de dispersión por nivel***:

- **Ninguno**. Con esta opción activa, el *Visor de resultados* no ofrece ni la prueba de Levene sobre homogeneidad de varianzas ni el gráfico de dispersión por nivel. Es la opción por defecto.

- **Estimación de potencia**. Cuando se incumple el supuesto de homogeneidad de varianzas (supuesto necesario para poder utilizar con garantía algunos procedimientos estadísticos como el *análisis de varianza*), es práctica frecuente aplicar algún tipo de transformación a los datos originales para conseguir homogeneizar las varianzas.

  Una transformación basada en *potencias* consiste en elevar las puntuaciones originales a una potencia específica. Para determinar la potencia apropiada, el SPSS genera un gráfico de dispersión comparando, para cada grupo, el logaritmo en base *e* de la mediana (en el eje de abscisas) con el logaritmo de la amplitud intercuartílica (en el eje de ordenadas). Cuando las varianzas son iguales, los puntos del gráfico se encuentran a la misma altura, es decir, alineados horizontalmente. La figura 11.12 muestra uno de estos gráficos referido a las variables *salario actual* (dispersión) y *nivel de estudios* (nivel).

  El gráfico de *Dispersión por nivel* de la figura 11.12 muestra 4 puntos, uno por cada nivel de la variable *estudios*. El hecho de que los puntos no se encuentren horizontalmente alineados indica que las varianzas no son homogéneas (lo cual coincide con la información proporcionada por el estadístico de Levene).

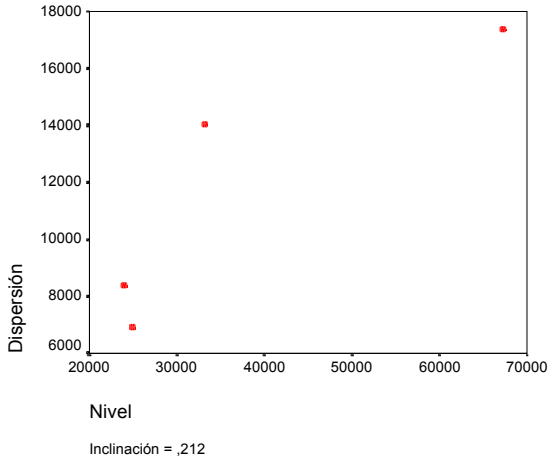
  El gráfico también muestra el valor de la pendiente de la recta de regresión (ver capítulo 18) obtenida por el método de mínimos cuadrados (*Inclinación* = 0,973). A partir del valor de la pendiente de la recta de regresión, el SPSS ofrece una estimación de la potencia a la que habría que elevar las puntuaciones de la variable *dependiente* (*salario actual*) para homogeneizar las varianzas de esa variable en cada nivel de la variable *factor*

(*nivel de estudios*); o, mejor, para intentar homogeneizarlas, porque lo cierto es que no siempre se consigue.

La estimación del valor de esa potencia se obtiene restando a uno el valor de la pendiente de la recta de regresión; en el ejemplo: 1 – 0,212 = 0,788. La potencia así estimada puede tomar cualquier valor. Sin embargo, lo habitual es utilizar potencias redondeadas a múltiplos de 0,5 (incluyendo el cero). Algunas de las potencias más utilizadas para transformar datos son las siguientes: –1 = recíproco; –½ = recíproco de la raíz cuadrada; 0 = logaritmo natural; ½ = raíz cuadrada; 1 = sin transformación; 2 = cuadrado; 3 = cubo. Todas estas transformaciones, que son las habitualmente recomendadas en la literatura estadística, están recogidas en la opción **Transformados**.

Con el valor de potencia obtenido en el ejemplo (0,788), utilizaríamos una potencia redondeada de 1, que equivale a no efectuar ningún tipo de transformación; lo cual significa que el SPSS no ha encontrado un valor de potencia que permita homogeneizar las varianzas.

Figura 11.12. Gráfico de dispersión por nivel de *salario actual* por *nivel de estudios*.



○ **Trasformados**. Una vez estimada la potencia apropiada para la homogeneización de las varianzas, podemos utilizar la opción **Transformados** para aplicar la transformación sugerida por el SPSS. Esta opción incluye, dentro de la lista desplegable **Potencia**, las siguientes transformaciones: logaritmo natural, recíproco de la raíz cuadrada, recíproco, raíz cuadrada, cuadrado y cubo. Todas estas transformaciones intentan homogeneizar las varianzas alterando (aumentando en unos casos y disminuyendo en otros) las varianzas de las distribuciones y corrigiendo el grado de asimetría.

La transformación *logarítmica* es apropiada para corregir distribuciones positivamente asimétricas. Y lo mismo vale decir de la transformación *raíz cuadrada* (si los valores de la variable transformada son pequeños, es conveniente sumar 0,5 a cada puntuación antes de obtener la raíz cuadrada). La transformación de los valores en sus *recíprocos* es adecuada cuando existen valores muy extremos por el lado positivo (cosa que ocurre, por ejem-

plo, con tiempos de reacción, donde los tiempos muy largos indican, probablemente, falta de atención más que otra cosa). Las transformaciones *cuadrado* y *cubo* permiten corregir, cada una en distinto grado, la asimetría negativa.

Al solicitar un gráfico de *Dispersión por nivel* seleccionando algún tipo de transformación, tanto la prueba de Levene como el gráfico de dispersión se obtienen a partir de los datos transformados. Pero, excepto en el caso de la transformación logarítmica, al solicitar una transformación basada en alguna de las potencias disponibles, el gráfico de dispersión por nivel se obtiene a partir de la mediana y de la amplitud intercuartílica, no a partir de sus logaritmos (estos valores, los logaritmos de la mediana y de la amplitud intercuartílica, son los que se utilizan en las opciones **Estimación de potencia** y **No transformados**).

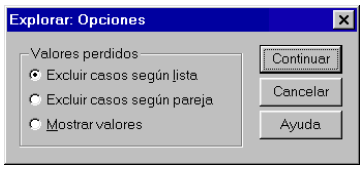
- **No transformados**. Esta opción permite obtener la prueba de Levene y el gráfico de dispersión por nivel a partir de los datos originales, sin ningún tipo de transformación (lo cual es equivalente a utilizar una potencia de 1). El gráfico de dispersión por nivel se obtiene comparando, para cada grupo, el logaritmo (en base *e*) de la mediana (en el eje de abscisas) con el logaritmo de la amplitud intercuartílica (en el eje de ordenadas).

Opciones

Las opciones del procedimiento **Explorar** permiten decidir qué tipo de tratamiento deseamos dar a los valores perdidos. Para ello:

- Pulsar el botón **Opciones...** del cuadro de diálogo *Explorar* (figura 11.1) para acceder al subcuadro de diálogo *Explorar: Opciones* que muestra la figura 11.13.

Figura 11.13. Subcuadro de diálogo *Explorar: Opciones*.



Valores perdidos. Las opciones de este recuadro permiten elegir una de las siguientes tres formas de tratar los valores perdidos:

- **Excluir casos según lista**. Se excluyen de todos los análisis solicitados los casos con algún valor perdido en cualquiera de las variables introducidas en las lista *dependientes* o en la lista *factores*. Es la opción por defecto.
- **Excluir casos según pareja**. Se excluyen de cada análisis concreto los casos con algún valor perdido en las variables que intervienen en ese análisis (no son excluidos de un análisis concreto los casos que, aun teniendo algún valor perdido en alguna variable de las listadas, no lo tienen en las variables objeto de ese análisis).

- **Mostrar valores**. Los casos con valores perdidos en la(s) variable(s) *factor* son tratados como una categoría más de esa(s) variable(s). Las tablas de frecuencias muestran los valores perdidos como una categoría más aunque esta opción no esté marcada.